L2/13-037R

Cedillas and commas below

Eric Muller, Adobe

January 29, 2013

The Marshallese orthography uses the Latin script, including the letters l, m, n, o with cedilla (with the cedilla attached to the rightmost leg of the m and n). It is considered unacceptable to display the cedillas as commas.

The Latvian orthography uses the Latin script, including the letters g, k, l, n, r with comma below. It is considered unacceptable to display the commas as cedillas.

Other orthographies use letters with commas below or cedilla (e.g. French, Romanian) and also require a specific form.

In principle, Unicode makes the distinction between cedillas and commas below, by providing U+0327 ◌̧ COMBINING CEDILLA and U+0326 ◌̦ COMBINING COMMA BELOW.

However, legacy character sets have not always made that distinction (at least in practice). For example, Latvian has historically used legacy ISO character sets that named the characters WITH CEDILLA, with the understanding that the context would imply the rendering appropriate for Latvian, i.e. with a comma below.

This ambiguity in legacy character sets has spilled over to Unicode, via the mapping from those legacy sets to Unicode. The mapping of the legacy WITH CEDILLA characters to the corresponding Unicode WITH CEDILLA characters was quite natural, but it carried with it the

ambiguity of the legacy characters. This can be seen in the code charts, which mentions that for example U+0146 ņ LATIN SMALL LETTER N WITH CEDILLA is to be used for Latvian and shows a representative glyph with a comma below.

Here are the code chart entries for the precomposed characters with cedilla:

00E7 ç LATIN SMALL LETTER C WITH CEDILLA
≡ 0063 c 0327 ◌̧

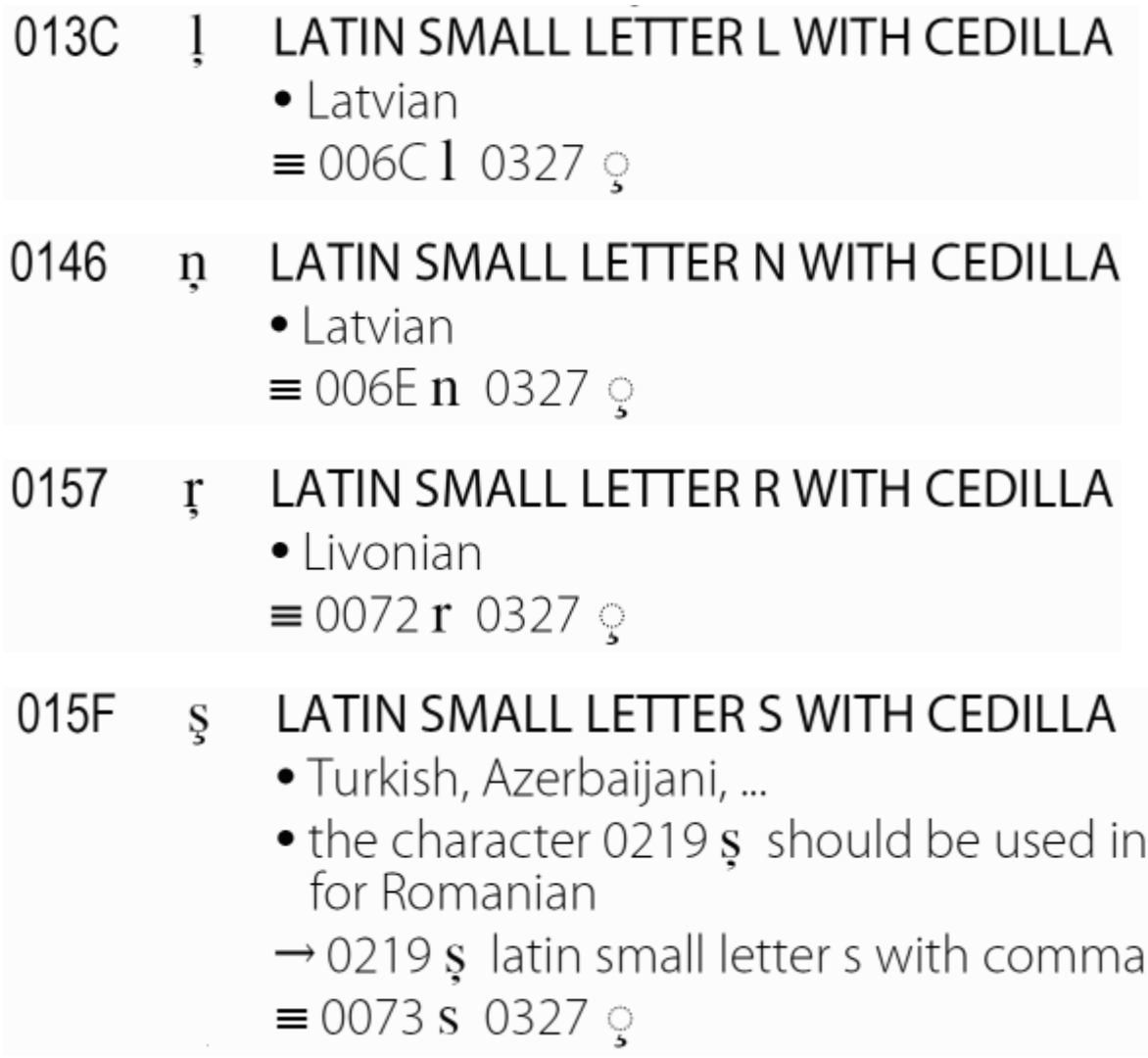
1E11 ḑ LATIN SMALL LETTER D WITH CEDILLA
• Livonian
≡ 0064 d 0327 ◌̧

0229 ȩ LATIN SMALL LETTER E WITH CEDILLA
≡ 0065 e 0327 ◌̧

0123 ģ LATIN SMALL LETTER G WITH CEDILLA
• Latvian
• there are three major glyph variants
≡ 0067 g 0327 ◌̧

1E29 ḩ LATIN SMALL LETTER H WITH CEDILLA
≡ 0068 h 0327 ◌̧

0137 ķ LATIN SMALL LETTER K WITH CEDILLA
• Latvian
≡ 006B k 0327 ◌̧

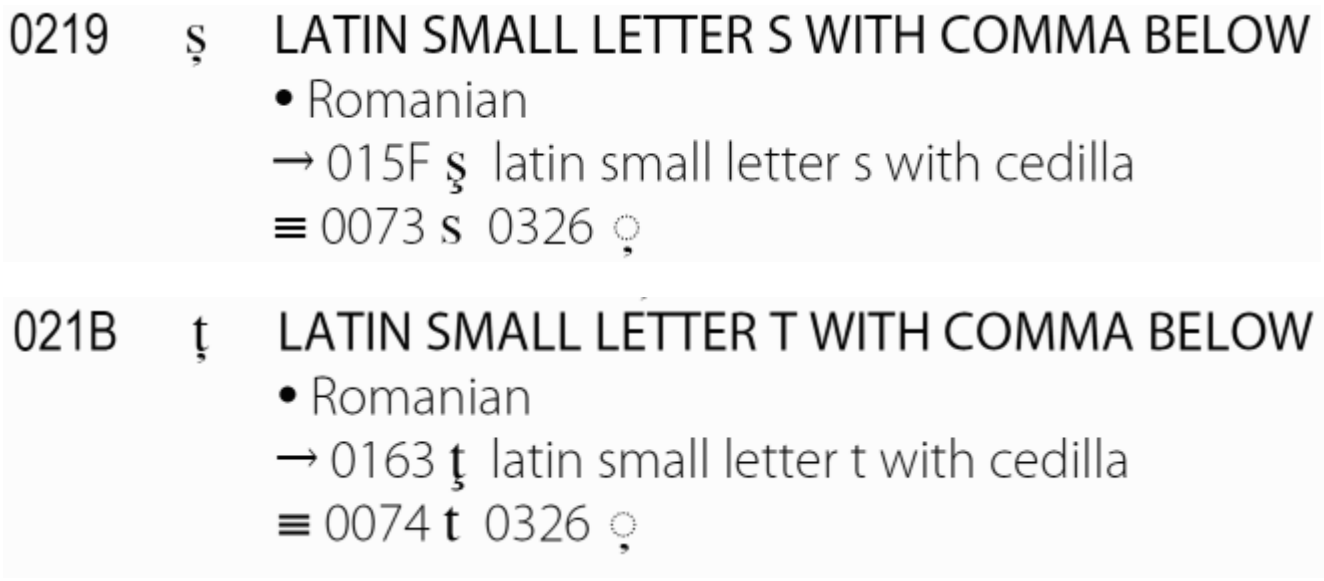


Note that the representative glyphs for c, e, h and s show a cedilla, the representative glyphs for d, k, l, n, r show a comma below, and the representative glyph for g show a turned comma above.

In Unicode 1.0 and 2.0, the representative glyph for N WITH CEDILLA showed a cedilla; this was changed in Unicode 3.0 to a comma below, to accomodate the use of the character in Latvian.

This situation creates a conflict: U+0146 ņ LATIN SMALL LETTER N WITH CEDILLA needs to be displayed with a comma for Latvian and with a cedilla for Marshallese. (Note that combining sequences vs. precomposed characters has no real importance, the same confusion happens with combining sequences.)

For completeness, here are the entries for the characters with comma below:



The same confusion was problematic for Romanian: the original recommendation for the representation of ș and ț was to use the LETTER WITH CEDILLA characters, and because those characters were often displayed with cedillas, the recommendation was changed to use the WITH COMMA BELOW characters.

It is clear that the confusion resulting from the current situation is problematic, and there are multiple paths the UTC could take. Needless to say, there is no easy solution, and concerned user communities should be involved in any case. Here is a (non-exhaustive) outline of possible approaches.

1) do nothing. This approach leaves the handling of the situation to implementers and users.

2) declare that comma below and cedilla are essentially two different renderings of the same abstract character. This approach has of course some serious problems: it does not match the real

world, in which users are quite attached to the difference; it is delicate architecturally, since we would have two distinct coded characters (U+0326 and U+0327) for the “same” abstract character; and may be more importantly, it destabilizes *all* the users of commas below and cedillas.

3) in an effort to mitigate the impact on existing data, leave the expected rendering of e.g. <U+0146 ņ LATIN SMALL LETTER N WITH CEDILLA> as a comma, and encourage <n, U+0327 ◌̧ COMBINING CEDILLA> to be rendered as a cedilla. This runs afoul of the canonical equivalence of those two sequences.

4) declare that comma below and cedilla are two different characters, and that rendering one by the other is not correct. This approach is of course problematic for communities which have used the WITH CEDILLA characters and want to see a comma, as it changes the representation of existing text. This affects mostly Latvian and Livonian.

I personally think that approach 4), however painful it may be, is the only one that has the potential of leading to reliable ecosystem. With a bit more details:

- recommend to use COMMA BELOW (combining or precomposed) when a comma is to be displayed

- recommend to use CEDILLA BELOW (combining or precomposed) when a cedilla is to be displayed

- change the representative glyph for the precomposed characters d/k/l/n/r WITH CEDILLA BELOW to show a cedilla

- replace the current annotation “Latvian” or “Livonian” on those characters by annotation similar to the one on U+015F ş LATIN SMALL LETTER S WITH CEDILLA: “the sequence <00xx, 0326> should be used instead for Latvian/Livonian”

- document the legacy situation and in particular the implications for mappings

- at least in rendering systems, it is possible to use locale information to render the appropriate form even for legacy representations

---